9/484,911

(12) UK Patent Application (19) GB (11) 2 314 178 (13) A

(43) Date of A Publication 17.12.1997

(21) Application No 9612261.9

(22) Date of Filing 12.06.1996

(51) INT CL[6]
G06F 17/30

(52) UK CL (Edition O)
G4A AUDB

(56) Documents Cited
EP 0304191 A2 WO 96/21901 A2
COMPUTER Record Access No 01929334 of Seybold Report on Desk top Publishing, v10, n8, p17(6), Apr 96

(58) Field of Search
UK CL (Edition O) G4A AUDB
INT CL[6] G06F 17/30
On-line: WPI, Inspec, Computer

(71) Applicant(s)
Infoseek Corporation

(Incorporated in USA - California)

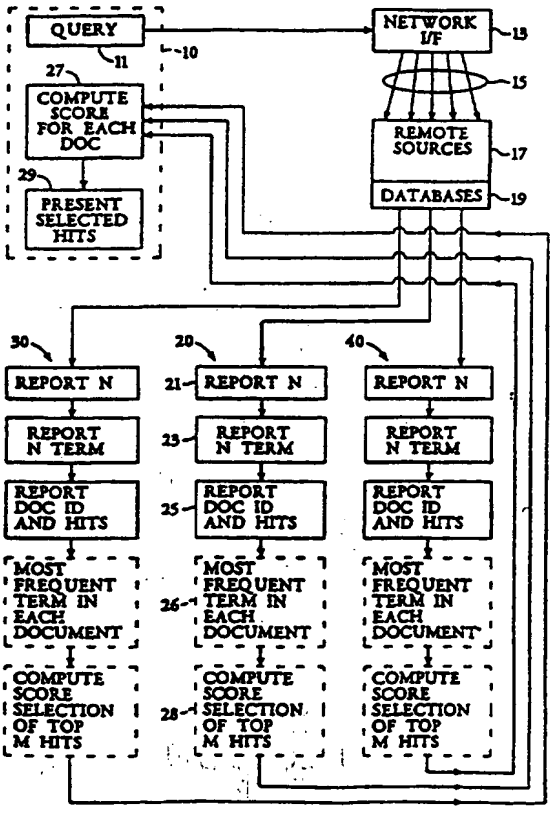
2620 Augustine Drive, No 250, Santa Clara, California 95054, United States of America

(72) Inventor(s)
Steven T Kirsch

(74) Agent and/or Address for Service
D Young & Co
21 New Fetter Lane, LONDON, EC4A 1DA, United Kingdom

(54) Searching multiple databases

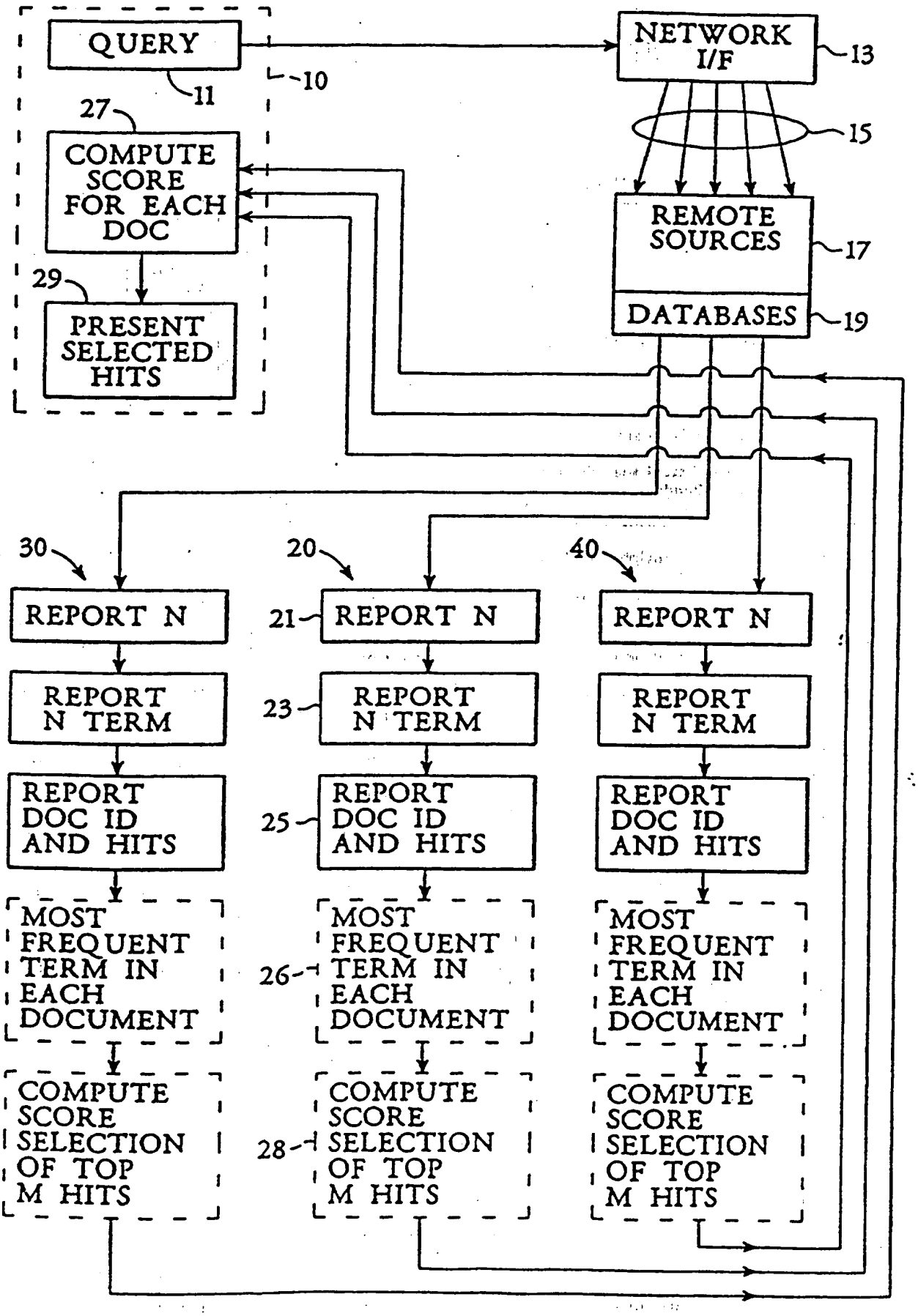
(57) A document search method using a plurality of databases 19 available from one or more servers 17 using one or more search engines. For each database, the number of records it holds is determined and reported 21, as well as the frequency of search query term occurrences or hits, 23, together with the identification of database records corresponding to the hits,25. Reports from a plurality of databases are furnished to a user terminal, a client,10, where client software computes, 27, a relevance score for each record based upon the number of records in the database, the number of records having at least one hit and the number of hits for each record. This local computation from uniform data allows all documents to be ranked consistently as if coming from a single database and regardless of differences in the search engines.



At least one drawing originally filed was informal and the print reproduced here is taken from a later filed formal copy.

This print takes account of replacement documents submitted after the date of filing to enable the application to comply with the formal requirements of the Patents Rules 1995

GB 2 314 178 A


QUERY
11
10
NETWORK I/F
13
15
27
COMPUTE SCORE FOR EACH DOC
REMOTE SOURCES
17
DATABASES
19
29
PRESENT SELECTED HITS
30
20
40
REPORT N
21
REPORT N
REPORT N
REPORT N TERM
23
REPORT N TERM
REPORT N TERM
REPORT DOC ID AND HITS
25
REPORT DOC ID AND HITS
REPORT DOC ID AND HITS
MOST FREQUENT TERM IN EACH DOCUMENT
26
MOST FREQUENT TERM IN EACH DOCUMENT
MOST FREQUENT TERM IN EACH DOCUMENT
COMPUTE SCORE SELECTION OF TOP M HITS
28
COMPUTE SCORE SELECTION OF TOP M HITS
COMPUTE SCORE SELECTION OF TOP M HITS

Description

Document Retrieval Over Networks

Technical Field

The invention relates to document searching and retrieval, particularly over networks.

Background Art

For more than twenty years, information services have provided access to multiple databases. For example, Dialog Information Services, now known as Knight-Ridder Information, Inc., provides several hundred databases (a.k.a collections) available to searchers. Some of these databases contain bibliographic abstracts, while others contain full-text documents. A searcher is able to apply a query to one or to a plurality of databases. At the outset, the searcher selects individual databases which are of interest, based on past experience, or selects a group of databases, selected by the information provider and related to a particular topic. For example, a searcher might select the topic of patents, a topic for which the information service has grouped a number of databases specific to patents. When a query is applied to the group of databases, the information service retrieves the number of hits in each database. The searcher then accesses databases of interest to view individual records. This system was originally designed for librarians and professional researchers who know where to look for desired information.

As wide area networks, such as Internet, become available, new opportunities in searching have become available, not only to searching professionals, but to lay users. New types of information providers are arising who use public, as well as private, databases to provide bibliographic research data and documents to users. When a user has an interest in a topic, such as

patents, he may not know what resources can be assembled for a search, nor the location of the resources. Since the resources frequently change, a searcher will have less interest in the source of the reply compared to the relevance of the reply. It has been recognized by others that distributed collections, available over wide area networks, can be treated as a single collection. Each sub-collection is searched individually, and the reports are combined in a single list. It has also been recognized by others that documents can be ranked by search engines in accord with an algorithm and given a weight, taking into account the nature of a particular collection. Document scores can be normalized to obtain scores that would be obtained if individual document collections were merged into a single, unified collection.

One of the problems that exists in the prior art is that the scores for each document are not absolute, but dependent on the statistics of each collection and on the algorithms associated with the search engines. A second problem which exists is that the standard prior art procedure requires two passes. In a first pass, statistics are collected from each search engine in order to compute the weight for each query term. In a second step, the information from the first step is passed back to each search engine, which then assigns a particular weight or score to each hit or identified document. A third problem that exists is that the prior art requires that all collections use the same search engine.

An object of the invention was to devise a method for searching multiple collections on a single pass, with ranking of documents on a consistent basis so that if the same document appears in two different databases, it would be scored the same when the results are merged. It is not required that the same search engine be used for all collections.

Summary of the Invention

The above object has been achieved with a document search and retrieval method which requires each participating search engine server to return statistics about each query term in each of the documents returned. A final relevance score is then computed at the client end, not the server. In this manner, all relevance scores are processed at the client in the same manner regardless of differences in the search engines.

Brief Description of the Drawing

Fig. 1 is a block diagram of the system of the present invention.

Best Mode for Carrying Out the Invention

With reference to Fig. 1, a query, indicated by the query block 11, is formulated by a user and applied to a terminal or client system. The query is electronically transmitted to a network interface 13. The network interface is an information service which has access to sources 17 having databases relating to the subject of the query. These databases, mounted on other servers, are simultaneously polled over communications channels 15, which may be wide area network links to the sources 17. The Internet is a model for such an arrangement of wide area network links and remote sources. The query is applied to search engines, represented by columns 20, 30, and 40, with each search engine accessing an associated database in block 19. Each search engine may have its own operating characteristics, such as Boolean logic, statistical inferences, etc. Each database produces a report containing the number of records, N, in the database, indicated by block 21. Also contained in the report is the number of times each search term occurs in documents responsive to the query. This quantity, N_{TERM}, is indicated by block 23. Thirdly, the report produces a document identification number for each document containing hits, together with the number of occurrences of each

search term, as indicated by block 25. From this information, a computation is made by the client software, using its own algorithm, of a score for each document, indicated by block 27. For example, a formula for computing a score is as follows:

$$\text{doc score} = \sum_{\text{all terms}} t_f \cdot idf$$

where t_f = number of occurrences of the term in the document and $idf = \log\left(\frac{N}{N_{TERM}}\right)$ where N and N_{TERM} are the sum of N and N_{TERM} values reported by all collections. The computed scores are transmitted to an output buffer, indicated by block 29, which sifts the top M scores from computation block 27, which have been requested by the person making the query. Note that only a single pass has been made through the database. Computed scores are treated as absolute values.

In an alternate embodiment, an optional parameter may be reported for use in the algorithm. Block 26 shows that the frequency of the most frequent term in each document is reported for purposes of using a more sophisticated ranking formula of the client. Another optional data reduction step is that each search engine may compute a score for document relevance in the manner known in the prior art. From this data, the search engine may preselect up to the top M hits in the database, where M is a maximum number of hits required by the user.

As an example, a search query might involve documents with the words "graphical user-interface". Table 1 below shows a report generated by a search engine which has selected a number of the highest ranking documents. This report is returned to the user's client software, where the user applies an algorithm, such as in formula (1) above, using term frequency data and document frequency data returned by each search engine. Thus, there is a local calculation of the document weighting for each query term, combining the N_{TERM} and N (= number of

documents) returned from each collection. Hence, term weighting is exactly the same as if the collections had been a single collection. Scoring is totally consistent even if different search engines participated in the search and the same document appearing in 2 different collections will always receive an identical score.

Table 1

N is 65,000 ← *total number of documents in the collection*

DocID	graphical	user-interface	
123	3	1	← *occurrences within the document of the term*
189	5	4	
100	4	2	
...	...	...	
232	32	2	
N_{TERM}	10000	23000	← *number of documents which contain the search term at least once*

Claims

1. A method for searching a plurality of databases which are distributed and accessible to a client through one or more search servers comprising,

(a) applying a search query from the client to each server associated with each database,

(b) obtaining, at the client from each server, statistics about each database,

(c) obtaining, at the client from each server, information about the documents resulting from application of the query to the database,

(d) computing, at the client, a score for each document using said statistics and said information, whereby the computed scores appear applicable to all databases as if the databases were joined into a single database.

2. The method of claim 1 wherein the statistics about said collection include the size of the collection in terms of the number of records.

3. The method of claim 1 wherein the information about each document includes the number of times each search term appeared in the document.

4. The method of claim 1 wherein the information about each database includes the number of documents containing each search term.

5. A method for searching a plurality of databases which are distributed and accessible to a client through one or more servers comprising,

(a) accessing each database from the client,

(b) applying a search query from the client to the server associated with each database,

(c) obtaining, at the client, statistics about each database,

(d) obtaining, at the client, statistical information about the relevant documents resulting from application of the query to the database,

(e) computing, at the client, a score for the relevant documents using said statistics and said information, whereby the computed score for a document is independent of the databasesit appears in.

6. A method for searching text documents among a plurality of databases, in response to a search query,

(a) applying a search query to each database,

(b) for each database determining the number of records,

(c) for each of said databases, applying the search query and recording the number of hits of each search query term and the identification of database records having at least one hit in said number of hits,

(d) for each of said databases, and for each query term, counting the records having at least one hit,

(e) reporting to a user a relevance score of each record with respect to the search query, computed using the results of steps (b), (c) and (d).

7. The method of claim 6 further defined by selecting a number of databases, having more than one search engine for the databases, prior to applying the same search query to all databases.

8. The method of claim 7 by selecting a number of records to be reviewed from among said number of databases, said number having the highest relevance scores for the search query.

9. The method of claim 8 further defined by preselecting a number of records prior to computing a relevance score.

The Patent Office

9

Application No: GB 9612261.9 **Examiner:** B G Western
Claims searched: 1 to 5 **Date of search:** 13 August 1996

Patents Act 1977
Search Report under Section 17

Databases searched:

UK Patent Office collections, including GB, EP, WO & US patent specifications, in:

UK Cl (Ed.O): G4A AUDB

Int Cl (Ed.6): G06F 17/30

Other: On-line : WPI, Inspec, Computer

Documents considered to be relevant:

Category	Identity of document and relevant passage	Relevant to claims
A	EP-0304191-A2 IBM See whole document	-
A	COMPUTER Record Access No. 01929334 of Seybold Report on Desktop Publishing, v10, n8, p17(6), 22 April 1996, "Searching far and wide: the powerful document retrieval software of PLS", Banet B.	-

X	Document indicating lack of novelty or inventive step	A	Document indicating technological background and/or state of the art.
Y	Document indicating lack of inventive step if combined with one or more other documents of same category.	P	Document published on or after the declared priority date but before the filing date of this invention.
&	Member of the same patent family	E	Patent document published on or after, but with priority date earlier than, the filing date of this application.

The Patent Office

10

Application No: GB 9612261.9
Claims searched: 6-9

Examiner: B G Western
Date of search: 30 June 1997

Patents Act 1977
Further Search Report under Section 17

Databases searched:

UK Patent Office collections, including GB, EP, WO & US patent specifications, in:

UK Cl (Ed.O): G4A AUDB

Int Cl (Ed.6): G06F 17/30

Other: On-line : WPI, Inspec, Computer

Documents considered to be relevant:

Category	Identity of document and relevant passage	Relevant to claims
A,E	WO-96/21901-A2 (PHILIPS) See whole document	-

X	Document indicating lack of novelty or inventive step	A	Document indicating technological background and/or state of the art.
Y	Document indicating lack of inventive step if combined with one or more other documents of same category.	P	Document published on or after the declared priority date but before the filing date of this invention.
&	Member of the same patent family	E	Patent document published on or after, but with priority date earlier than, the filing date of this application.